título

Procesamiento masivo de datos mediante Cassandra y Spark

Máster en Sistemas Informáticos Avanzados
Septiembre de 2016

Autor:

Xabier Zabala Barandiaran

Supervisores:

German Rigau i Claramunt
UPV/EHU
Iñigo Etxabe y Beñat Aranburu
Datik Información Inteligente S.L.

Agradecimientos

En primer lugar, quisiera expresar mi gratitud hacia las personas que han posibilitado la concepción y el desarrollo de este proyecto. Agradezco a German Rigau i Claramunt, supervisor del proyecto por parte de la UPV/EHU, la predisposición mostrada y el asesoramiento ofrecido durante el transcurso del mismo. Doy también las gracias a Inigo Etxabe y Beñat Aranburu, supervisores del proyecto por parte de Datik Información Inteligente S.L., por poner a mi disposición todos los medios tecnológicos necesarios para llevar a cabo el proyecto y por el trato ofrecido desde el primer día.

Agradecerles, cómo no, a mis padres José Javier Zabala y María Pilar Barandiaran por el esfuerzo desempeñado para facilitarme, en la medida que les ha sido posible, el camino que he recorrido hasta llegar aquí. Me congratula haber sabido responder satisfactoriamente a la confianza que ellos siempre han depositado en mí. Por todo lo que han hecho y por todo lo que suponen para mí, un beso enorme para los dos.

No quisiera olvidarme de todas las personas que han estado a mi lado durante este maravilloso periplo, a los cuales no me atrevo a mencionar de forma individual por el miedo de dejar a alguno en el tintero. Gracias a los amigos de toda la vida por el apoyo ofrecido durante este camino. Gracias a los compañeros de la facultad, por todo los momentos vividos juntos y en especial a aquellos que durante esta etapa se han ganado a pulso el privilegio a ser parte importante de lo que me resta de existencia.

Por último, pero no por ello menos importante, quisiera evocar a todos los docentes que han tomado parte en mi formación desde aquel Septiembre del 2008 y agradecer a todos ellos el conocimiento compartido y el esfuerzo invertido en mí durante estos años.

Gracias de todo corazón a la gente mencionada en este breve capítulo por haber hecho de mí un mejor profesional y sobre todo una mejor persona.

Resumen

Proyecto Final del Máster en Sistemas Informáticos Avanzados. Estudio de carácter empírico realizado sobre el rendimiento ofrecido por varias tecnologías emergentes en el campo del Big Data en comparación a una base de datos tradicional a la hora de operar en escenarios que requieren un almacenamiento y procesamiento eficaz de volúmenes masivos de datos.

Para llevar a cabo el experimento, dos entornos de prueba totalmente aislados han sido erigidos sobre la misma máquina física. En el primero, se ha configurado un clúster compuesto por cuatro nodos virtuales que operan dentro de una red privada. Dichos nodos han sido dotados de tecnología necesaria para el funcionamiento de Apache Cassandra [**apachecassandra**] y Apache Spark [**apachespark**]. En el segundo, se ha instalado una base de datos MySQL tradicional sobre un único nodo virtual que hereda la potencia total de la máquina física. Una vez habiendo poblado las bases de datos mediante un data-set público de aproximadamente 25GB y diseñado unas consultas acorde a la naturaleza de los datos, se han ejecutado dichas consultas para así cuantificar el tiempo de respuesta que necesitan en cada escenario.

El estudio evidencia que a la hora de trabajar con volúmenes masivos de datos el binomio entre Apache Cassandra y Apache Spark mejora sustancialmente los tiempos de procesado obtenidos con MySQL además de ofrecer una solución totalmente escalable. No obstante, para gozar de las ventajas que ofrecen estas nuevas tecnologías, se antoja necesario un análisis previo de los datos a tratar, aspecto en el que MySQL ofrece una mayor libertad.

Palabras Clave: Big Data, Apache Cassandra, Apache Spark, MySQL, Comparativa.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Desde Aristóteles y su libro Segundos Analíticos hasta Galileo, padre de la ciencia moderna, muchos adalides del conocimiento han proclamado que un método de investigación basado en lo empírico y en la medición, sujeto a los principios específicos de las pruebas de razonamiento es el camino para alcanzar la verdad.

Hoy en día, época en la que los avances tecnológico han posibilitado observar y medir de forma exhaustiva un gran abanico de fenómenos, la ingente cantidad de datos que se genera en el proceso es, a veces, intratable por medio de las tecnologías convencionales, y por ende, imposible extraer conocimiento de ellos. El problema, lejos de atenuarse, se acrecienta con el paso del tiempo, ya que, estudios como el realizado por McKinsey Global Institute (MGI) estiman que el volumen de datos que se genera está creciendo un 40 % cada año y auguran que entre 2009 y 2020 se verá multiplicado por 44[1].

Por ello, en los últimos años ha irrumpido la necesidad de encontrar metodologías y herramientas que permitan procesar y extraer el conocimiento que atesora el torrente de información en la cual se encuentra envuelta la sociedad, dando como resultado el nacimiento del Big Data.

El mundo empresarial, por su parte, no se ha mantenido al margen de esta gran revolución. Conscientes de los beneficios que les puede reportar en diferentes aspectos como en el análisis de mercado y calidad de los servicios que ofertan, la gran mayoría de las empresas se han interesado en el Big Data. De un estudio realizado entre los altos ejecutivos de las firmas que lideran el Wall Street se desprende que el 96 % tiene planeadas ciertas iniciativas relacionadas con el Big Data, y el 80 % tiene finalizada alguna[2].

1.1. Contexto

Datik Información Inteligente S.L. es una empresa tecnológica perteneciente al Grupo Irizar que desarrolla soluciones ITS destinadas a la gestión del trasporte, tanto ferroviario como por carretera y movilidad ciudadana.

Uno de los productos estrella de la entidad es el denominado iPanel, concentrador de información que ofrece al operador de transporte servicios de valor añadido en la gestión de la información generada por su flota. El funcionamiento de este servicio se puede resumir mediante la Figura 1:

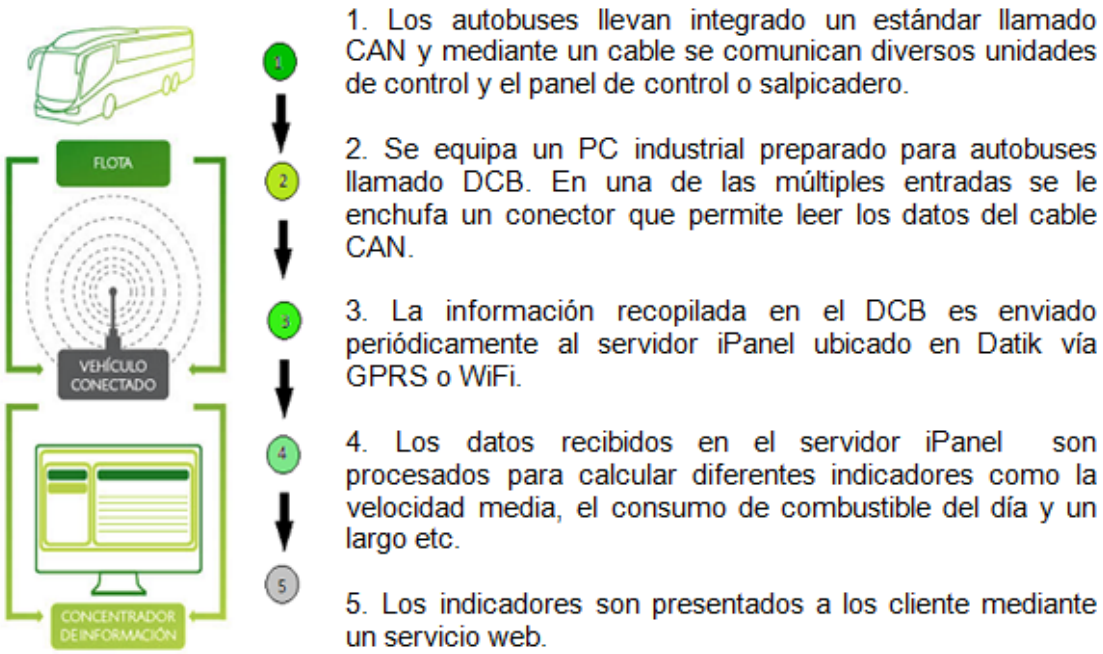

Figura 1: Funcionamiento resumido de iPanel

El incesante aumento en el número de vehículos equipados genera un crecimiento exponencial de los datos que se han de almacenar y procesar en una base de datos MySQL. Aunque a día de hoy la situación se encuentra bajo control y no conlleve peligro alguno para el funcionamiento del servicio, Datik tiene identificados varios peligros que en un futuro cercano podrían comprometer dicho funcionamiento.

El primero de todos, es el tiempo necesario para realizar el cálculo de los indicadores. Se trata de un proceso ejecutado una vez al día que atendiendo los datos recopilados en

las últimas 24 hora vuelve a calcular todos los indicadores de iPanel. Ello implica realizar operaciones aritméticas sobre diferentes campos como la velocidad y el consumo de combustible y agrupar los resultados por cliente, flota, vehículo o un espacio temporal. Actualmente, se necesitan varias horas para finalizar la computación, pero, debido al aumento de los datos a tratar, es posible que en un futuro existan graves dificultades para realizar el calculo en menos de 24 horas, invalidando así la funcionalidad de ofrecer los indicadores del último día.

Otro de los problemas, intrínseco al uso de una base de datos centralizada, es el operar sobre un único punto de fallo. Debido a que la mayoría de procesos pasan por dicho punto, existe un alto riesgo de sufrir el denominado efecto dominó, esto es, que la caída provocada por un servicio acarree la del resto. Actualmente, Datik dispone un servidor de réplica capacitado para suplir al primario en caso de que ocurra algo así. No obstante, esta práctica condenar una máquina al ostracismo ya que sus recursos quedan desaprovechados en el 99,95 % del tiempo.

EL último, es la corrupción de datos. Es un fenómeno que puede suceder (y sucede) debido a un bug, fallo de almacenamiento inesperado, o una caída de MySQL cuando el resultado del checksum de una página es diferente al esperado. Este fenómeno podría comprometer seriamente los datos que Datik ofrece a sus clientes mediante la aplicación web, hecho que se h de evitar a toda costa.

1.2. Propuesta

1.3. Organización del documento

En esta memoria se ha documentado el desarrollo de la herramienta ***exerClick***, dentro del Trabajo de Fin de Grado (TFG) del autor. En el documento se describe la propuesta, la planificación y gestión que esta lleva consigo, la implementación llevada a cabo y las conclusiones finales.

En este primer capítulo se ha introducido el problema a resolver y se ha explicado la propuesta presentada en este proyecto.

En el capítulo 2 se presenta el Documento de Objetivos de Proyecto (DOP). Este recoge el alcance y las fases y tareas del proyecto, el análisis de riesgos y el análisis de factibilidad.

Una vez en el capítulo 3 se explica la gestión llevada a cabo durante el proyecto. Se presentan las metodologías utilizadas: Metodologías Ágiles e InterMod (adaptada a las necesidades de este proyecto). A continuación se detallan cada una de las iteraciones llevadas a cabo (como parte de la metodología InterMod): duración, objetivos y tareas realizadas. Al final del capítulo se muestra la documentación asociada a las iteraciones y los objetivos, además del seguimiento de tiempo realizado.

A continuación, en el capítulo 4 se detalla el análisis de requisitos. Primero se detallan los requisitos no-funcionales y luego los funcionales (prototipos en papel llevados a cabo durante las primeras iteraciones que dan una visión global del proyecto).

En el capítulo 5 se explica el diseño e implementación llevados a cabo. Se comienza mostrando la estructura de documentos del proyecto, luego el diseño realizado en base al análisis de requisitos del capítulo 4 y finalmente una visión general de la implementación de la lógica de negocio.

Para finalizar, en el capítulo 6 se presentan las conclusiones, líneas futuras para el proyecto y las lecciones aprendidas.

Fuera de la estructura general de la memoria, tenemos la bibliografia y los apéndices. En estos últimos tenemos las actas de reuniones, las actas de pruebas y la vista de relaciones de la base de datos (de la parte utilizada o creada específicamente para el proyecto).